CBMI2022

StyleGAN-based CLIP-guided Image Shape Manipulation

Yuchen Qian, Kohei Yamamoto, Keiji Yanai

The University of Electro-Communications, Tokyo, Japan

BACKGROUND

- Image transformation and image editing based on deep learning are being studied.
- Natural language is becoming an interface between humans and machines.

⇒Natural language image editing using multimodal models is attracting attention.

Open-Edit (ECCV 2020)

Xihui Liu, Zhe Lin, Jianming Zhang, Handong Zhao, Quan Tran, Xiaogang Wang and Hongsheng Li. Open-Edit: Open-Domain Image Manipulation with Open-Vocabulary Instructions. ECCV, 2020

StyleCLIP (ICCV 2021)

Or Patashniky, Zongze Wu, Eli Shechtmanx, Daniel Cohen-Ory and Dani Lischinskiz. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. arxiv: 2103.17249

PROBLEM

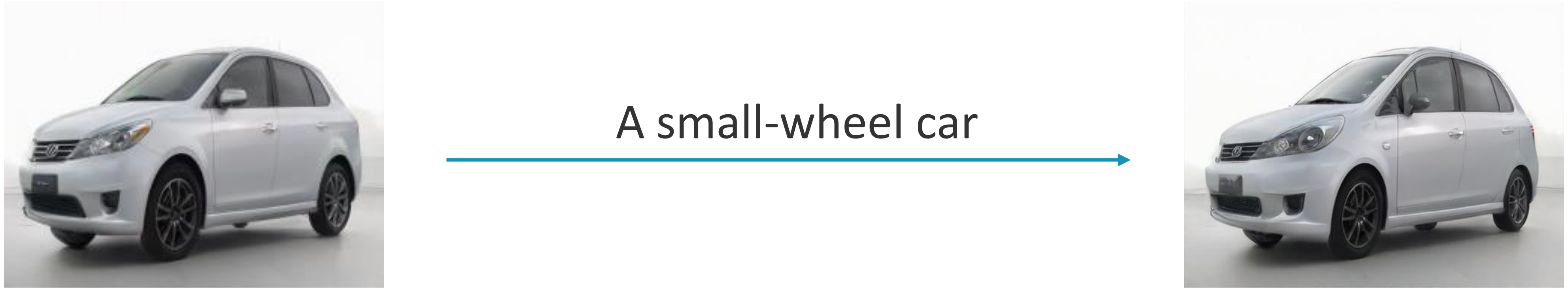
- Prior research has focused on editing for appearance features. (e.g., color and texture)
- In contrast, there are few studies on editing for shape features. (e.g., some sizes)

ManiGAN (CVPR 2020)

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz and Philip HS Torr. Manigan: Text-guided image manipulation. CVPR, 2020

AIM

- To achieve editing of image shape features based on input text using pre-trained GAN models and multimodal models



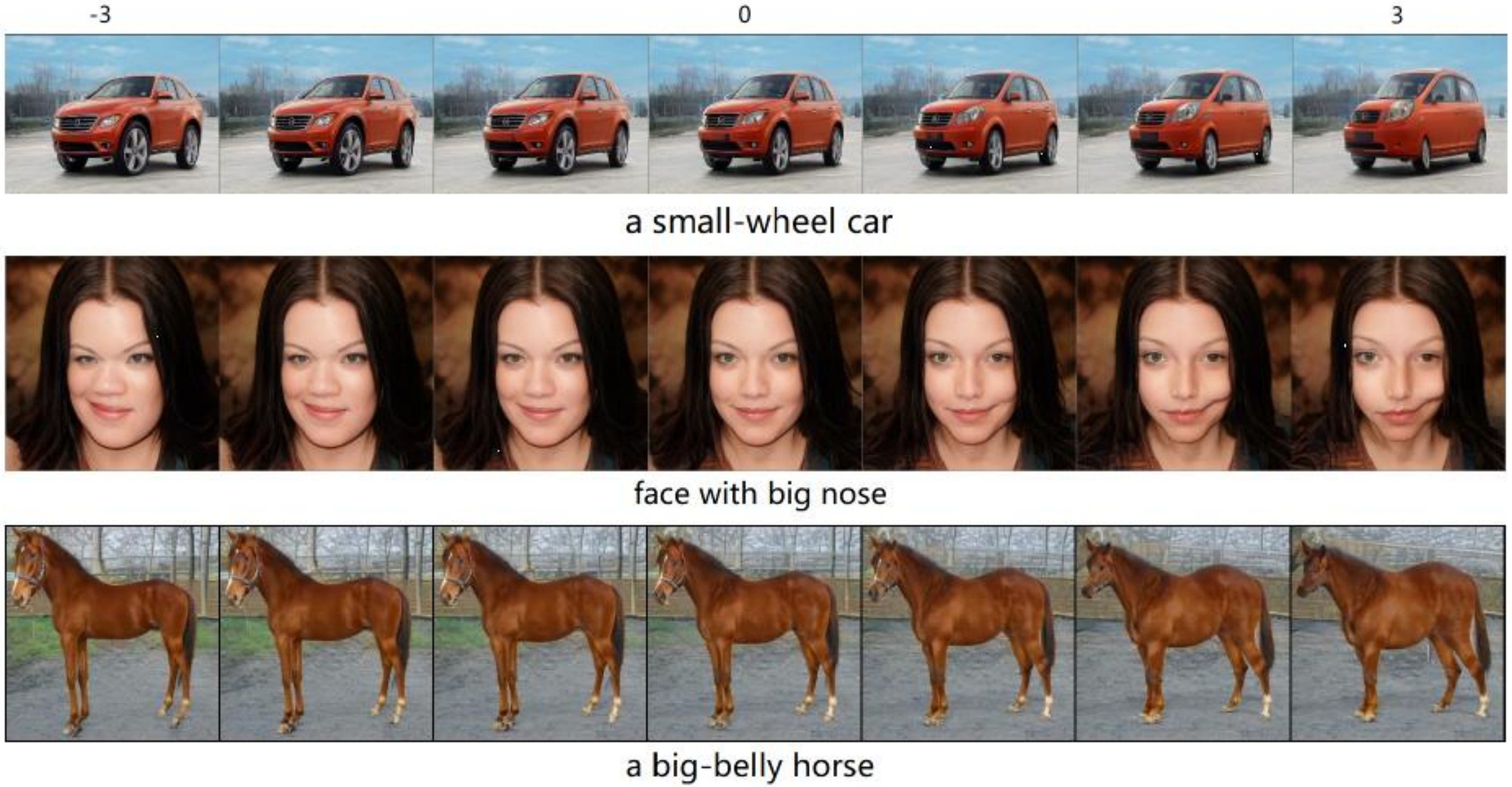
APPROACH

- Based on NaviGAN's idea of tuning generator parameters, the model was built using the pre-trained StyleGAN2.
- CLIP measures the similarity between the generated images and the input text and provides loss for optimization.

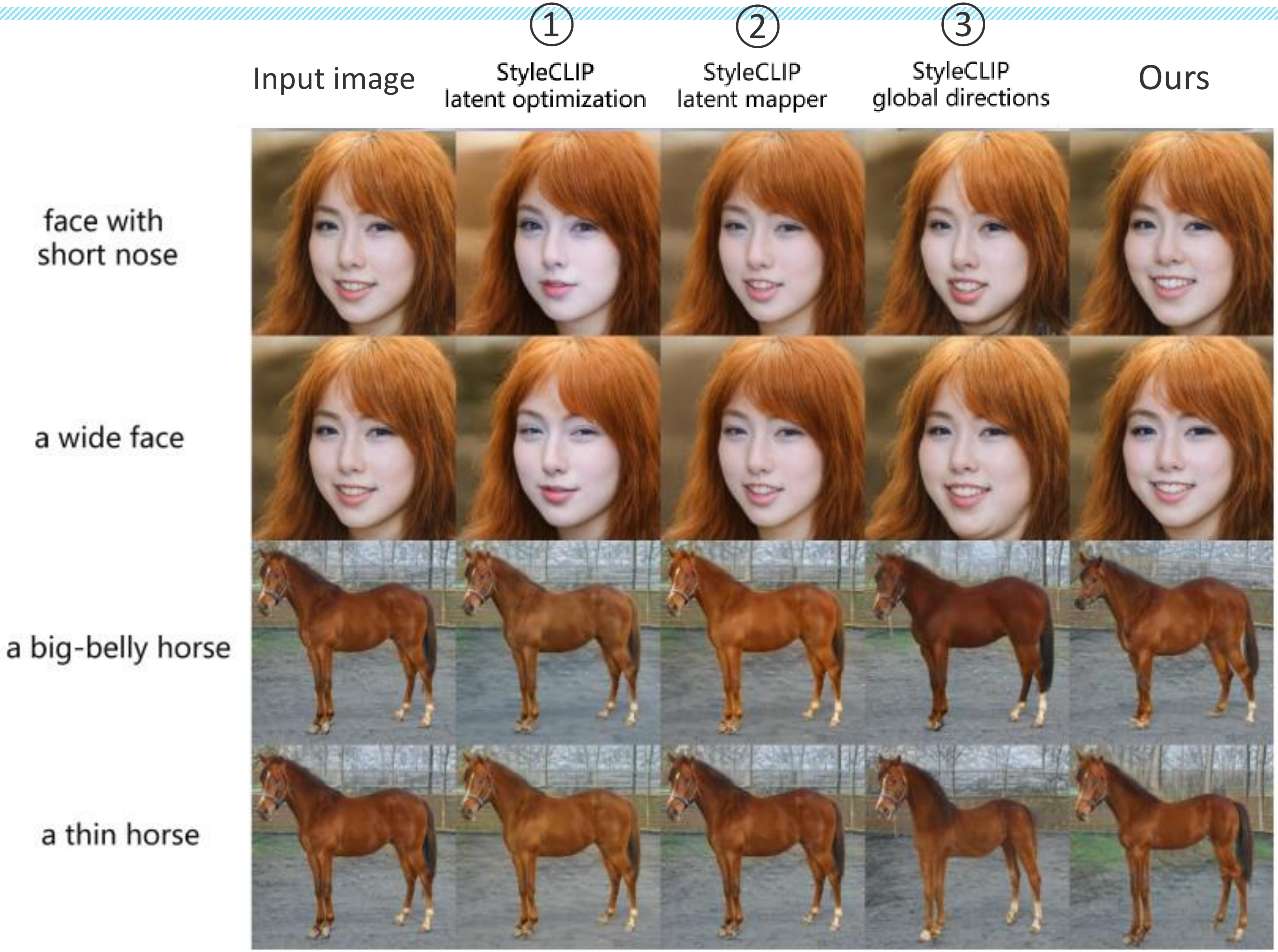
A small-wheel car

QUALITATIVE RESULTS

- Edited images produced by shift x multiplied by a factor of -3~3



QUALITATIVE RESULTS : Comparisons



QUANTITATIVE RESULTS

- Evaluation by the index called FID
- Calculate FID values with 3000 real images and generated images.
- In many cases, the proposed method obtained better FID values.
- The proposed model can keep the image quality better than StyleCLIP.

	magnitude	-10	-5	-3	0	+3	+5	+10
Wheel Size	StyleCLIP	54.34	42.35	33.36	-	18.16	23.33	67.57
	Ours	26.27	17.96	15.34	-	15.22	21.30	62.39
	GAN inversion	-	-	-	12.54	-	-	-
Cheek Size	StyleCLIP	30.55	29.16	28.90	-	23.40	29.20	30.12
	Ours	30.21	29.34	28.89	-	27.96	28.41	29.86
	GAN inversion	-	-	-	25.6	-	-	-

CONCLUTION

- Proposed method edits shape features of images based on input text.
- The proposed method's qualitative and quantitative representation in editing shape features outperforms the conventional method of adjusting the latent space.

- Only one target feature can be edited in one optimization.
  - Learning and optimization methods capable of editing multiple features
- Large changes in features other than the target feature
  - Minimize impact on other features as much as possible.